DeepNews: Detecting Quality in News

Eun Seo Jo, Aashiq Muhamed, Shashank Nuthakki, Ayush Singhania

{eunseo, aashiq, nuthakki, ayushs}@stanford.edu
Advisor: Frederic Filloux

Introduction

Problem: Classifying high (value-added) and low (commodity) quality news articles.
Approach: Use POStagged articles as data to a Deep learning model for classification with emphasis on sequence learning.

Data

News Provider	#Articles (1000s)	Label
Guardian	112	High
Economic Times	91	High
Reuters	800	High
Financial Times	1000	High
Quartz	800	High
Huffington Post	116	Low
Business Insider	129	Low
CRN	122	Low

Table 1: News Data Collection

Data Collection: Data scrapped from the URLs given by the providers.
Labeling: As human labeling is expensive, we used the reputation of the news provider as the indicator of each article's quality.

Part of Speech Tagging

Instead of using the raw English text to train the model we used the POStagged text.
Why?
Focus on syntactic structure of Language

- Masking all topical words to counter biases and difference across news collections and providers.
- Helps in generalizing to unseen providers (Fig. 4.)
- Easier to train ☺

Fig. 1: Data Processing: POS tagging

Models

Pre-training: We used Glove 100d word embeddings as word representations for all of our models
Baseline: The baseline is nearest centroid prediction, with 100d vector article representations using CBOW model on POSTagged text.
Feedforward Network with CBOW Inputs
Input: 100d vector representation of articles from doing CBOW, with high/low quality labels.
Model: We used a 3 hidden layer network, with number of hidden units for the layers being 1000, 500, 100, followed by a final sigmoid layer.
LSTM based Recurrent Neural Network
Input: Each example has 50 sequence tokens and with high/low quality source label.
Model: We used a 2-layer LSTM network with 500 and 100 hidden units and 50 time-steps. The final time-step output form layer 2 is followed by a 3 layer feedforward network.

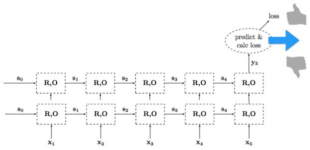

Fig. 2: Figure of LSTM-RNN

Language Model for Value-added News
Input: Each example has 50 sequence tokens.
Model: We used a 1-layer LSTM network, with 500 hidden units to learn a sequence of length 50. The trained LM is used to get text encodings which are fed to a feedforward discriminator.

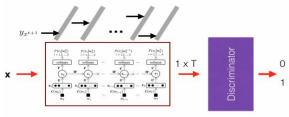

Fig. 3: LM Classifier Model

Results

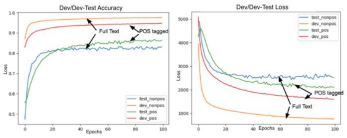

Fig. 4: POS/nonPOS Diff on Feedforward

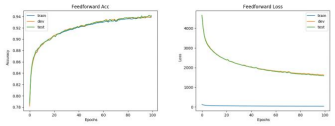

Fig. 5: Feedforward Acc/Loss

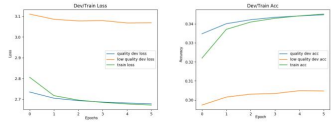

Fig. 6: LM loss/acc on Train and High/Low Quality Dev

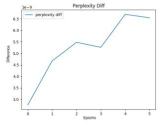

Fig. 7: LM Perplexity Differences between High/Low Quality

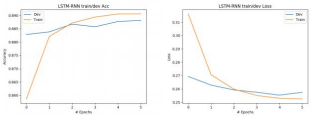

Fig. 8: LSTM-RNN model loss/acc

Conclusions

From these experiments we discovered several notable findings.

- The feedforward CBOW model performs well.
- The sequence models performed much better when combined with a feedforward layer.
- The style difference between quality and commodity news was significant enough that all of our models were able to meet relatively high accuracy even with POS masking.

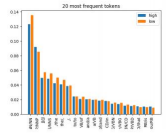

Fig. 9: 20 Most Freq Token Types High/Low Quality

- Model Comparison

Model	Test Accuracy
Classification by Centroid	70.32
Feedfoward CBOW Model	93.42
LSTM Model	96.26
Language Model Classifier	In-progress

Table 2: Performance Comparison across Models

Future Directions

- Computational constraints limited training the sequence models on the entire dataset as well as using much bigger models
- Scoring (1-5 scale) of articles based on the perplexity scores obtained from the trained Language model is a promising direction for this project.

References

[1] Jozefowicz, et al. Exploring the Limits of Language Modeling. 2016
[2] Arisoy, Ebru, et al. Bidirectional Recurrent Neural Network Language Models for Automatic Speech Recognition. 2015.
[3] Bengio, Yoshua, et al. A Neural Probabilistic Language Model. 2003